CAS751: IT Metods in Trustworthy ML Homework 4 - **Due: 08/12/2024**

Differential Privacy and Machine Learning

Instructor: Shahab Asoodeh

1. (30 points) Let the dataset $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be given, where x_i and y_i for $i \in \{1, 2, \dots, n\}$ represent feature vectors and labels, respectively. Consider the typical minimization problem

$$\min_{\theta} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i)),$$

where $\ell(\theta, (x_i, y_i))$ quantifies the loss associated with representing datapoint (x_i, y_i) with the model parameter $\theta \in \mathbb{R}^d$.

- Describe gradient descent and stochastic gradient descent (SGD) algorithms for solving this minimization problem.
- Let q_D^t denote the query at the t^{th} iteration of SGD algorithm. Considering the gradient clipping constant C, we can describe q_D^t as:

$$q_D^t = \begin{cases} n\nabla\ell(\theta_{t-1}, (x_i, y_i)) & \text{if } \|\nabla\ell(\theta_{t-1}, (x_i, y_i))\|_1 \le C \\ nC\frac{\nabla\ell(\theta_{t-1}, (x_i, y_i))}{\|\nabla\ell(\theta_{t-1}, (x_i, y_i))\|_1} & \text{if } \|\nabla\ell(\theta_{t-1}, (x_i, y_i))\|_1 > C. \end{cases}$$

  Notice that this is the ℓ_1-norm clipping (as opposed to ℓ_2-norm clipping discussed in class). Follow the three steps given in lecture to characterize the privacy guarantee of SGD algorithm when we add Laplace noise in each iteration.

2. (40 points.) Given a closed bounded (i.e., compact) set $\mathcal{C} \subset \mathbb{R}^d$, define the projection operator $\Pi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ by

$$\Pi_{\mathcal{C}}(a) = \arg\min_{b \in \mathcal{C}} \|a - b\|_2.$$

With this definition at hand, let's define the following variant of SGD algorithm (known as projected noisy SGD, or PNSGD): Let the dataset $\mathbb{D} = \{x_1, \dots, x_n\}$ and an arbitrary distribution μ_0 on $\mathcal{C}$ be given. The algorithm initiates with $Y_0 \sim \mu_0$ and iterates as follows:

$$Y_t = \Pi_{\mathcal{C}}(Y_{t-1} - \eta[\nabla\ell(Y_{t-1}, x_t) + N_t]),$$

to generate $Y_1, Y_2, \dots, Y_n$, where $N_t \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Assume that the loss function $\ell(\cdot, x)$ is L-Lipschitz for any x. This algorithm is detailed in the following:

Algorithm 1 PNSGD Algorithm

Require: Dataset $\mathbb{D} = \{x_1, \dots, x_n\}$, learning rate $\eta > 0$, initial point $Y_0 \sim \mu_0$ and iid copies N_t of $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$
  for $t \in \{1, \dots, n\}$ **do**
    $Y_t = \Pi_{\mathcal{C}}(Y_{t-1} - \eta[\nabla\ell(Y_{t-1}, x_t) + N_t])$
  end for
  return Y_n

There are two main differences between this algorithm and the one we saw in the class:

- Here, the algorithm goes over the dataset sequentially with batch size = 1. That is, each datapoint contributes exactly once. Therefore, the number of iteration is equal to the size of the dataset; $T = n$.

- All intermediate parameters $Y_1, Y_2, \ldots, Y_{n-1}$ are assumed to be hidden. The algorithm releases only Y_n.

Notice that the t^{th} iteration of this algorithm can be described as the composition of three operations: 1. the function $\psi_t(y) \coloneqq y - \eta \nabla \ell(y, x_t)$, 2. Gaussian noise addition with variance $\mathcal{N}(0, \eta^2 \sigma^2 \mathrm{I}_d)$, and 3. projection operator $\Pi_{\mathcal{C}}$. Thus, the PNSGD algorithm can be viewed as the concatenation of n channels $\mathsf{K}_1, \ldots, \mathsf{K}_n$ as depicted in Fig. 1.

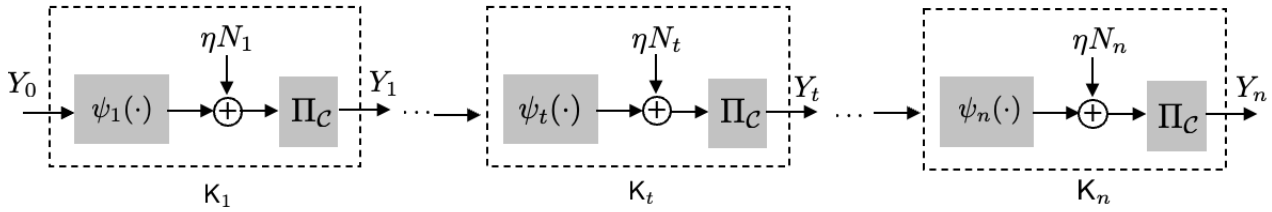

Figure 1: PNSGD algorithm running on dataset $\mathbb{D}$

Let μ_t be the distribution of Y_t (or equivalently the output distribution of K_t, which is in fact the input distribution of K_{t+1}).

(a) Consider a neighboring dataset $\mathbb{D}' = \{x_1', x_2, \ldots, x_n\}$. Similar as before, PNSGD algorithm for this dataset can be viewed as the concatenation of n channels $\mathsf{K}_1', \mathsf{K}_2 \ldots, \mathsf{K}_n$ as depicted in Fig. 2, where $\psi_1'(y) \coloneqq y - \eta \nabla \ell(y, x_1')$.

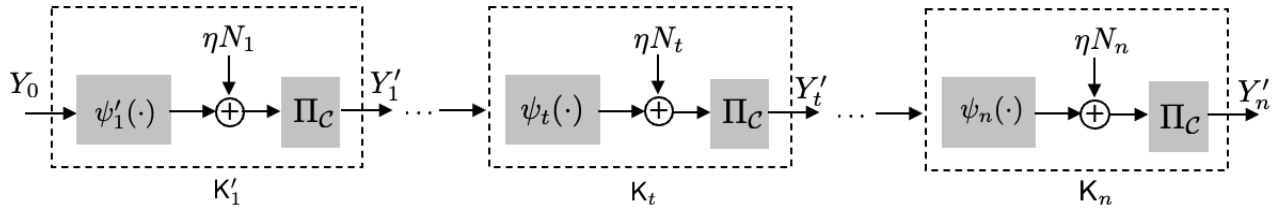

Figure 2: PNSGD algorithm running on dataset $\mathbb{D}'$

Since only Y_n is released, differential privacy guarantee is determined by deriving $\mathsf{E}_{e^\varepsilon}(\mu_n \| \mu_n')$ where μ_n' is the distribution of Y_n'. Use DPI and the contraction coefficient under hockey-stick divergence to upper bound $\mathsf{E}_{e^\varepsilon}(\mu_n \| \mu_n')$.

**You may need the following*: If the loss function $y \mapsto \ell(y, x)$ is L-Lipschitz for any x, then it can be shown that $||\psi_i(y_1) - \psi_i(y_2)||_2 \leq \|\mathcal{C}\|_2 + \eta L$ for any $y_1, y_2 \in \mathcal{C}$, where $\|\mathcal{C}\|_2$ is the diameter of $\mathcal{C}$. (No need to prove it, but feel free to do so!)*

(b) Now consider another neighboring dataset $\mathbb{D}' = \{x_1, x_2, \ldots, x_{n-2}, x_{n-1}', x_n\}$. Again use DPI and the contraction coefficient under hockey-stick divergence to upper bound $\mathsf{E}_{e^\varepsilon}(\mu_n \| \mu_n')$.

(c) (**Bonus. 15pt**) As you can see in you answers to Parts (a) and (b), $\mathsf{E}_{e^\varepsilon}(\mu_n \| \mu_n')$ depends on the index in which $\mathbb{D}$ and $\mathbb{D}'$ differ. As such, different individuals in the dataset are promised different levels of privacy (which is not what you expect from an appropriate privacy-preserving mechanism!). How would you modify the algorithm to resolve this issue?

3. (30 points.)

(a) Prove that for any pair of distributions P and Q and any $\gamma \geq 1$, we have $\mathsf{E}_\gamma(P \| Q) \geq \gamma \mathsf{TV}(P, Q) + 1 - \gamma$.

(b) Let M be an ε-LDP mechanism, i.e., $\mathsf{E}_{e^\varepsilon}(\mathsf{M}_x \| \mathsf{M}_{x'}) = 0$ for any possible input x and x', where M_x is the output distribution of the mechanism M when the input is x. Prove that $\mathsf{TV}(\mathsf{M}_x, \mathsf{M}_{x'}) \leq 1 - e^{-\varepsilon}$.

(c) Prove that

$$\mathsf{M} \text{ is } \varepsilon\text{-LDP} \iff \eta_{e^\varepsilon}(\mathsf{M}) = 0.$$

(Recall that $\eta_\gamma(\mathsf{M})$ denotes the contraction coefficient of M under the hockey-stick divergence.)

(d) Prove that

$$\mathsf{M} \text{ is } \varepsilon\text{-LDP} \implies \eta_{\mathsf{TV}}(\mathsf{M}) \leq 1 - e^{-\varepsilon}.$$